2nd Place Solution to Google Landmark Retrieval 2021

Yuqi Zhang, Xianzhe Xu, Weihua Chen, Yaohua Wang, Fangyi Zhang, Fan Wang, Hao Li
Alibaba Group

{gongyou.zyq, xianzhe.xxz, kugang.cwh, xiachen.wyh, zhiyuan.zfy,
fan.w, lihao.lh}@alibaba-inc.com

Abstract

This paper presents the 2nd place solution to the Google Landmark Retrieval 2021 Competition on Kaggle. The solution is based on a baseline with training tricks from person re-identification, a continent-aware sampling strategy is presented to select training images according to their country tags and a Landmark-Country aware reranking is proposed for the retrieval task. With these contributions, we achieve 0.52995 mAP@100 on private leaderboard. Code available at https://github.com/WesleyZhang1991/Google_Landmark_Retrieval_2021_2nd_Place_Solution

1. Introduction

Image retrieval is a very important computer vision task which aims at finding images similar to the query image. It is different from instance-level retrieval. Image retrieval aims to retrieve objects holding the same appearance with the query, even they are not the same instance. This makes the task more easier compared to instance-level retrieval. On the other hand, it is different from the fine-grained level image retrieval. The fine-grained level image retrieval pays more attention on the local attentions to discover more details due to its small intra-category variance, such as person re-identification.

Landmark retrieval is an instance-level retrieval task, which aims to search the same landmark from a large candidate set. In this paper, we will introduce our techniques used in the fourth landmark retrieval competition, Google Landmark Retrieval 2021 [1] held on Kaggle. Some of them are inspired from the state of the art algorithms in person re-identification.

Besides, we also involve many techniques that are commonly used in previous competitions [10, 12, 14, 8, 2], including model structures [16], training strategies, loss functions [7, 6, 5, 20]. These techniques have been well explored and introduced in previous competitions, so we only introduce our methods, denoted as new contributions listed below.

- We involve bags-of-tricks from person re-identification and conduct careful experiments on these tricks.

- We propose a continent-aware sampler to balance the distribution of training images based on their continent tags.

- We design a Landmark-Country aware reranking algorithm and integrate it with the K-reciprocal reranking method.

2. Method

2.1. Baseline network

We select several large CNN networks including SE-ResNet-101 [9], ResNeXt-101 [18], ResNeSt101 and ResNeSt269 [19] as backbones. IBN extension [15] is used for SE-ResNet and ResNeXt-101. The input size is selected as 384 for pretraining and 512 for the last fine-tuning. The last stride of the CNN network is set to 1. We use generalized mean-pooling (GeM) for pooling method with p=3.0. Arcface loss [7] with $scale = 30$ and $margin = 0.3$ is used. We use weight decay=0.0005. Training details with gradually enlarging input size and data scale can be found in Section 3.3.

Some tricks from person re-identification has been explored. Random Erasing [22] randomly erases out image patches and has shown great success in many fields. Label smoothing [13] by using soft targets that are a weighted average of the hard targets can often be useful in many computer vision tasks.

2.2. Sampling Strategy

On one hand, in person re-identification or face recognition, id-uniform is widely used as a data sampling strategy. For a batch, we randomly select P Ids and then K images for each Id. Thus we have P*K images as a batch. Each id is treated fair for this setting. On the other hand, softmax

sampling has been widely used in previous competitions. The softmax sampling just shuffle all dataset once at the beginning of the epoch and then samples iterative through the data. Head data which appears more will be put more attention with this setting. We have tried these sampling strategies and find neither of them are good enough for our task.

As the provided landmark dataset pays more concentration on Asia landmarks [11], we manage to design a sampling strategy based on their continent labels. First, we use the country-and-continent-codes-list [3] to find how many countries each continent has. Then we search and list all the landmarks in each country from [4]. Based on this processing, we can get the country tag and continent tag for every landmark.

We setup a continent sampling prob by 'Asia': 0.5, 'Europe': 0.2, 'Africa': 0.15, 'North America': 0.1, 'South America': 0.02, 'Antarctica': 0.01, 'Oceania': 0.01, 'OTHER': 0.01. For an epoch of images, we sample continent images by the corresponding ratios. Also, we learn from previous paper [10] which set 0.66 probability for clean data and 0.33 probability for noisy data.

2.3. Reranking

Reranking is very essential to the final performance. Besides K-reciprocal reranking [21], a Landmark-Country aware reranking algorithm is specially designed for our task.

We observe that there are many images which contain the same landmark and can't be easily retrieved by visual features due to great variations caused by views and illumination. Considering this, a Landmark-Country aware reranking is proposed by taking fully use of the training set. Specifically, as each image in training set has its landmark tag and its country tag, we first assign query and index images in the testing set with a training tag, and then retrieval the query from index images by the assigned tags.

For query image tagging, We give a list of potential landmark tags and country tags to every query image according to its top K similar images in training set, and each landmark tag and country tag is scored by the accumulation of similarity in top k, as shown in Equation (1). The jth tag appears in topk is noted as $score_j$. For each index image, the landmark tag and country tag are given by its most similar image in training set.

$$score_j = \sum_i^k sim_{i==j} \qquad (1)$$

Then we reranking the query's retrieval results using Equation (2).

$$sim_{final} = sim + \alpha \times L_{score} + \beta \times C_{score} \qquad (2)$$

Trainset	# Samples	# Labels
clean	1,580,470	81,313
c2x	3,223,078	81,313
full	4,132,914	203,094
all	4,825,830	203,094

Table 1. Training set construction.

where α and β is the weight of the landmark tag score and the country tag score respectively.

With the Landmark-Country aware reranking, images have the same landmark tag or country tag with the potential tag of query are advanced in the retrieval sorting.

3. Experiments

3.1. Training set construction

The official GLDv2 dataset [17] has provided a clean and a full version. The clean version include clean data, while the full version contain both clean and noise data. As pointed out by previous works [2], many noisy images in the full set with the labels of clean set have been filtered out during the cleaning stage. Expanding these noisy data into clean set results in 'c2x'. Although the existing of noise, these noisy images may contain very valuable information, e.g., the indoor or outdoor of a building. What's more, we also include the index set from GLDv2, which shares many common ids with the full set. We list the dataset in Table 1 for a better understanding.

3.2. Validation Set

We use the 1129 queries in GLDv2 test set together with the 76,176 index set from the competition. We put all GT images of each query into the index set and expand it to 78,959 images. In this case, any query could find all its GTs in the index set.

3.3. Implementation details

Motivated by previous solutions, we first train the models on 'clean' subset with an input size of 384(For ResNeSt269, the size is 448). The initial learning rate is 0.01 and we train for 10 epochs with the first epoch as warmup. Then we keep the same category number but with more data as 'c2x'. The initial learning rate is 0.001 and we train for 6 epochs. Then we expand dataset to 'full' and train for another 2 epochs with the initial learning rate as 0.0001 and input size as 512. At last, we use 'all' data from GLDv2 and train for another 2 epochs with the initial learning rate as 0.0001 and input size as 512.

3.4. Bag-of-tricks

Many tricks from person re-identification have been tested and we list some important ones in Table 2. Since

Setting	Validation	Public	Private
baseline	32.60	28.44	30.59
+RE	32.78	29.29	30.60
+label smooth	32.55	28.21	29.79

Table 2. Effect of different tricks on landmark retrieval.

Setting	Validation	Public	Private
id-uniform	31.05	24.78	27.28
softmax	32.60	28.44	30.59
continent-aware	33.07	31.37	32.44

Table 3. Effect of different sampling strategy.

the dataset is very large, we use R50 backbone with an input size of 256×256. We list the validation accuracy as well as public/private scores. Ransom erasing proves to be effective while label smoothing fails.

3.5. Sampling strategy

Table 3 lists results for different sampling strategy. The widely used id-uniform strategy fails for landmark retrieval. We think the reason is due to the large amount of tail data. These tail data may be noisy and thus drifts the model training. The softmax strategy works better than id-uniform. For continent-aware strategy, it achieves the best performance.

4. Conclusion

In this paper, we import techniques and bag-of-tricks from person re-identification. For the specific task of landmark retrieval, we propose continent-aware sampling strategy and Landmark-Country aware post processing, which has proven to be very effective on the private leaderboard. The relationship between landmark retrieval and landmark recognition would be studied in the future.

References

[1] https://www.kaggle.com/c/landmark-retrieval-2021.

[2] https://storage.googleapis.com/kaggle-forum-message-attachments/1035342/17203/2nd_Place_Solution_to_Google_Landmark_Recognition_Competition_2020_bestfitting.pdf.

[3] https://datahub.io/JohnSnowLabs/country-and-continent-codes-list.

[4] https://storage.googleapis.com/gld-v2/web/index.html.

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017.

[6] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3988–3994, 2017.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

[8] Christof Henkel and Philipp Singer. Supporting large-scale image recognition with out-of-domain samples. *arXiv preprint arXiv:2010.01650*, 2020.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[10] SeungKee Jeon. 1st place solution to google landmark retrieval 2020. *arXiv preprint arXiv:2009.05132*, 2020.

[11] Zu Kim, André Araujo, Bingyi Cao, Cam Askew, Jack Sim, Mike Green, N Yilla, and Tobias Weyand. Towards a fairer landmark recognition dataset. *arXiv preprint arXiv:2108.08874*, 2021.

[12] Ke Mei, Jinchang Xu, Yanhua Cheng, Yugeng Lin, et al. 3rd place solution to" google landmark retrieval 2020". *arXiv preprint arXiv:2008.10480*, 2020.

[13] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

[14] Kohei Ozaki and Shuhei Yokoo. Large-scale landmark retrieval/recognition under a noisy and diverse dataset. *arXiv preprint arXiv:1906.04087*, 2019.

[15] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018.

[16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[17] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

[18] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[19] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[20] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang. Cross-view gait recognition by discriminative feature learning. *IEEE Transactions on Image Processing*, 29:1001–1015, 2019.

[21] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017.

[22] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 34, pages 13001–13008, 2020.